Remote Sensing Image Super-Resolution for Imbalanced Textures: A Texture-Aware Diffusion Framework

Enzhuo Zhang Sijie Zhao Dilxat Muhtar Zhenshi Li Xueliang Zhang† Pengfeng Xiao
Nanjing University
zenzhuo@smail.nju.edu.cn, zxl@nju.edu.cn

6. Summary

The Supplementary Material includes the following:

- We provide additional methodological details, including supplementary explanations of the RTDM estimation in section 7.1 and the detailed architecture of the RTDM prediction network in section 7.2.
- We provide additional experimental details and results, including the experimental setup in section 8.1, comparisons with officially released checkpoints in section 8.2, comparison with Inpainting-like Baselines in section 8.3, additional ablation studies in section 8.4, and a discussion of the limitations of no-reference metrics in section 8.5.
- We discuss the limitation of this work in section 9.
- We provide additional visual comparison results, including results from our retrained model and those obtained with the official weights of the baseline methods.

7. Methodological Details

7.1. RTDM Estimation

In the following, we introduce two key components of RTDM estimation: the Contrast Consistency Term (CCT) and Spatial LPIPS.

Contrast Consistency Term (CCT). Building on the contrast component of SSIM [20], CCT compares the local statistics of two images—the preliminary super-resolution result generated by a PSNR-oriented model I^{PSR} and its ground truth I^{HR}—to expose high-frequency deficiencies, thus enabling CCT to identify local texture discrepancies. For each pixel, we first compute the local standard deviation:

$$\sigma(i,j) = \mathrm{sd}\big(R\big(i-\tfrac{m-1}{2} : i+\tfrac{m-1}{2},\, j-\tfrac{m-1}{2} : j+\tfrac{m-1}{2}\big)\big), \tag{7}$$

where $\sigma(i,j)$ is the local standard deviation at (i,j), $sd(\cdot)$ represents the local standard deviation operator, R denotes the local window, and m specifies its size, which is fixed at 11.

Then, following the contrast term of SSIM, the pixel-wise contrast-consistency map is computed as:

$$M_{CCT}(i,j) = \frac{2\,\sigma_{PSR}\sigma_{HR} + C}{\sigma_{PSR}^2 + \sigma_{HR}^2 + C}, \tag{8}$$

where σ_{PSR} and σ_{HR} are the local standard deviations of I^{PSR} and I^{HR}, respectively, C is a small stabilising constant.

Spatial LPIPS. Spatial LPIPS is a spatially variant extension of Learned Perceptual Image Patch Similarity (LPIPS) [24]. Whereas the original LPIPS collapses the perceptual distance between two images into a single scalar, spatial LPIPS preserves spatial information by producing a full-resolution error map that assigns a perceptual dissimilarity score to every pixel. Each image in the pair is independently passed through a pretrained network (e.g, VGG [16], AlexNet [7]), and the feature maps of N convolutional layers are extracted. For each layer l, we compute the channel-wise squared difference between the $\ell_2 - normalized$ feature maps of the two images, I^{PSR} and I^{HR}:

$$d_l(i,j) = \sum_{c=1}^{C_l} w_{l,c} \left(\tilde{f}^{PSR}_{l;c,i,j} - \tilde{f}^{HR}_{l;c,i,j} \right)^2 \tag{9}$$

where $d_l(i,j)$ is the spatial perceptual error at coordinate (i,j) in layer l; C_l is the number of channels in that layer; $w_{l,c}$ is the LPIPS-learned weight for channel c of layer l; $\tilde{f}^{PSR}_{l;c,i,j}$ and $\tilde{f}^{HR}_{l;c,i,j}$ are the corresponding $\ell_2 - normalized$ activation values.

Since several intermediate feature maps have a lower spatial resolution $(H_l \times W_l)$ than the input image $(H \times W)$, they must be upsampled before aggregation:

$$\bar{d}_l = \mathrm{Upsample}\big(d_l,\, (H,W)\big), \qquad \bar{d}_l \in \mathbb{R}^{H\times W}. \tag{10}$$

The final spatial LPIPS map is obtained by averaging the upsampled maps across all N selected layers:

$$M_{SL}(i,j) = \frac{1}{N} \sum_{l=1}^{N} \bar{d}_l(i,j). \tag{11}$$

†Corresponding author.

7.2. RTDM Prediction Network

In the following, we provide a detailed description of the RTDM prediction network architecture.

As illustrated in Figure 6, the RTDM prediction network adopts a dual-encoder, single-decoder architecture. The LR encoder takes an LR image of spatial size $H_L \times W_L$ as input, while the SR encoder accepts the paired preliminary SR (PSR) image generated by a PSNR-oriented model, with spatial resolution $4H_L \times 4W_L$. The LR encoder branch first encodes the LR image with a series of convolutional layers and then upsamples the features using a PixelShuffle operation, producing a feature map of size $(N, C, 2H_L, 2W_L)$. The SR encoder branch symmetrically encodes the PSR image using convolutional layers followed by a PixelUnshuffle operation that downsamples the spatial resolution, yielding a feature map of size $(N, 4C, 2H_L, 2W_L)$. Both branches are equipped with residual blocks and squeeze-and-excitation (SE) blocks [5] to enhance local representation and perform channel-wise recalibration.

The LR and SR features are then projected into the same channel space and fused by a pixel-wise gated fusion module, where a learned gate map adaptively modulates the relative contribution of LR and SR features at each spatial location. The fused features are fed into a U-Net-style decoder composed of convolutional layers and residual blocks. Finally, a PixelShuffle layer followed by a convolutional prediction head reconstructs a single-channel output at full PSR resolution $(N, 1, 4H_L, 4W_L)$, producing the predicted RTDM.

8. Experiments

8.1. Experimental Details

Implementation Details of Main Experiment. Our TexADiff framework comprises three constituent models: the PSNR-oriented model (SwinIR [8]), the RTDM prediction network, and the texture-aware denoising diffusion model. All models were trained on 4 NVIDIA A800 GPUs and evaluated on NVIDIA RTX 3090 GPUs. The remaining training details are provided below:

- **SwinIR (PSNR-oriented model).** We train the model on a randomly selected subset (13,630 images) of the training dataset, cropping each image into 256 × 256 patches and using a batch size of 48 for 200k optimization steps. Training uses the Adam [6] optimizer with an initial learning rate of 2×10^{-4}; the learning rate is halved at steps 50k, 75k, and 100k.
- **RTDM prediction network.** We train it on the full training set, cropping images into 512 × 512 patches with a batch size of 48 for 30k steps. Training uses the AdamW [12] optimizer with an initial learning rate of 1×10^{-4} and the learning rate is updated according to a cosine annealing schedule [11].
- **Texture-aware denoising diffusion model.** Our model is based on SDXL-base [15]. We train it on the full training set, pairing each training image with a caption, either directly from RS5M [25] or generated by Qwen2.5-VL [1] when no caption is available. Training is carried out for 30k optimization steps on 512 × 512 image patches cropped from each image, using a batch size of 256. We use AdamW optimizer and assign separate initial learning rates of 5×10^{-5} for the U-Net and 5×10^{-6} for the MiniControlNet, both decayed by a cosine-annealing schedule.

At inference, the classifier-free guidance [4] scale is fixed to 5 and the total number of sampling steps is fixed to 20.. Unless otherwise specified, the text prompt is set to null.

For other models that also leverage T2I pre-trained text-to-image priors, we train them with the same hyperparameters used for our method. Importantly, for FaithDiff [2] we follow the original paper's protocol and perform an additional 6k pre-training steps. When retraining ResShiftL [23], we crop all input images to 256 × 256, use a batch size of 96, and optimize for 200k steps. When retraining Real-ESRGAN [19], all input images are cropped to 256 × 256. The model is first optimized for 100k steps with a batch size of 96 without adversarial loss, and then further optimized for 50k steps with a batch size of 48 with adversarial loss.

Implementation Details of Ablation Experiments. Due to limited computational resources, the ablation experiments are configured slightly differently from the main comparative experiments. For ablations that only modify the inference procedure—namely (i) impact of different RTDM at inference, and (ii) impact of textual descriptions—we do not retrain any models and instead reuse the checkpoints from the main comparisons. All other ablation studies are trained for 10,000 steps with a batch size of 96. Unless the corresponding component or strategy is being ablated, the TDAL weight α is fixed to 1, the RTDM binarization threshold is set to 0.4 at inference, the sampling interval is fixed to $[100, 500]$, and the text prompt is left empty.

8.2. Comparison with Official Checkpoints

In the comparisons of main text, we retrain all competing methods on the same remote sensing image dataset to ensure a fair evaluation. Table 10 reports the results obtained using their officially released checkpoints, together with the differences (Δ) relative to our retrained versions. Retraining yields consistent improvements across all evaluation metrics for every diffusion-based model. Compared with the results based on the official checkpoints, our method further enlarges its lead on perceptual metrics (LPIPS [24] and DISTS [3]). Furthermore, we compare our method with

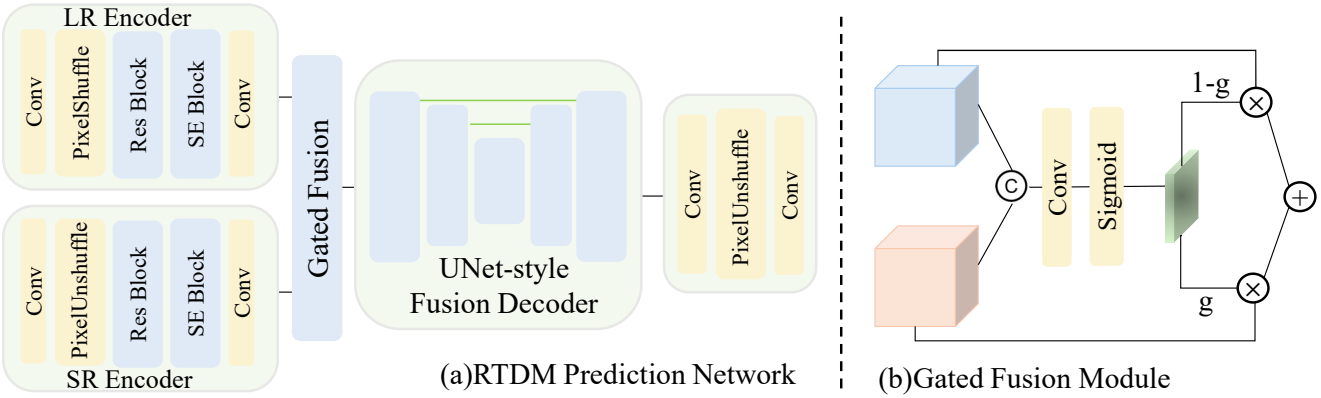

Figure 6. Detailed RTDM prediction network architecture.

Table 10. Quantitative comparison using officially released checkpoints of comparison methods on the AID, DOTA-Test, LoveDA-Val and RSC11 datasets. Best results are in **bold**, second-best are underlined. Δ denotes the change in the metric relative to our retrained version (metric Δ = metric(official) - metric(retrain)). ↑ indicates that higher is better. ↓ indicates that lower is better.

Datasets	Methods	PSNR↑	PSNR Δ	SSIM↑	SSIM Δ	LPIPS ↓	LPIPS Δ	DISTS↓	DISTS Δ
AID	Real-ESRGAN	**23.24**	0.02	**0.5231**	-0.0049	0.4309	0.0039	0.2344	-0.0065
	ResShiftL	21.88	-1.15	0.3861	-0.0643	0.5595	0.1762	0.2989	0.0385
	PASD	22.67	-0.53	0.4915	-0.0146	0.4541	0.0328	0.2366	0.0300
	FaithDiff	22.03	-0.76	0.4358	-0.0338	0.4127	0.0375	0.2344	0.0324
	SUPIR	21.11	-	0.3792	-	0.5030	-	0.2653	-
	HYPIR	20.46	-	0.3935	-	0.4643	-	0.2443	-
	Ours(thr=0.35)	22.87	-	0.4696	-	**0.3788**	-	**0.1939**	-
	Ours(thr=0.40)	22.62	-	0.4554	-	0.3823	-	0.1953	-
LoveDA-Val	Real-ESRGAN	25.56	-1.32	**0.6032**	-0.0481	0.4497	0.0466	0.2312	0.0068
	ResShiftL	24.13	-1.35	0.4266	-0.0861	0.6398	0.1406	0.2977	0.0491
	PASD	25.19	-0.84	0.5579	-0.0239	0.4921	0.1371	0.2179	0.0585
	FaithDiff	23.79	-2.54	0.4191	-0.1600	0.5273	0.1700	0.2748	0.0875
	SUPIR	24.23	-	0.4536	-	0.5783	-	0.2426	-
	HYPIR	22.11	-	0.4385	-	0.5507	-	0.2511	-
	Ours(thr=0.35)	**26.34**	-	0.5779	-	0.3264	-	0.1766	-
	Ours(thr=0.40)	26.15	-	0.5662	-	**0.3253**	-	**0.1751**	-
DOTA-Test	Real-ESRGAN	**25.58**	0.10	**0.6288**	-0.0010	0.4236	0.0435	0.2691	0.0333
	ResShiftL	23.93	-1.06	0.4774	-0.0778	0.5678	0.1163	0.3193	0.0371
	PASD	24.55	-0.53	0.5871	-0.0128	0.4656	0.0501	0.2692	0.0381
	FaithDiff	23.64	-0.98	0.5260	-0.0348	0.4410	0.0962	0.2646	0.0516
	SUPIR	22.89	-	0.4421	-	0.5544	0	0.3160	-
	HYPIR	22.91	-	0.5166	-	0.4529	-	0.2600	-
	Ours(thr=0.35)	24.74	-	0.5711	-	**0.3358**	-	**0.1934**	-
	Ours(thr=0.40)	24.52	-	0.5580	-	0.3417	-	0.1978	-
RSC11	Real-ESRGAN	**21.71**	0.14	**0.4337**	0.017	0.5349	-0.0179	0.3027	0.0202
	ResShiftL	21.28	-0.29	0.3519	-0.0204	0.5953	0.0685	0.3325	0.0311
	PASD	21.27	-0.35	0.4051	-0.0035	0.6255	0.1254	0.3166	0.0420
	FaithDiff	20.42	-0.23	0.3358	-0.0158	0.5104	0.0243	0.2895	0.0309
	SUPIR	19.65	-	0.2724	-	0.6507	-	0.3516	-
	HYPIR	19.87	-	0.3370	-	0.5188	-	0.2832	-
	Ours(thr=0.35)	20.80	-	0.3574	-	0.4708	-	0.2374	-
	Ours(thr=0.40)	20.56	-	0.3460	-	**0.4693**	-	**0.2364**	-

Table 11. Comparison with mask-guided baseline.

Method	PSNR↑	LPIPS↓	DISTS↓
mask-guided	22.65	0.4003	0.2048
Ours	**22.78**	**0.3883**	**0.2038**

Table 12. Ablation of TADL adjustable weight.

Weight	PSNR ↑	LPIPS ↓	DISTS ↓
0	22.59	0.4001	0.2039
0.5	22.59	0.3967	0.2043
1	**22.78**	**0.3883**	**0.2038**

Table 13. Ablation of texture-aware sampling schedules over different intervals.

Interval	PSNR ↑	LPIPS ↓	DISTS ↓
[0,500]	22.62	0.3938	**0.1990**
[100,400]	22.77	0.3902	0.2043
[100,500]	22.78	**0.3883**	0.2038
[100,600]	22.78	0.3905	0.2044
[100,700]	22.78	0.3906	0.2043
[200,500]	22.78	0.3910	0.2045

Table 14. Ablation of different textual descriptions.

Text Prompt	PSNR ↑	LPIPS ↓	DISTS ↓
Null Prompt	22.62	0.3823	0.1953
MLLM-generated Prompt	22.56	0.3886	0.1961

SUPIR [22] and HYPIR [9], both of which are trained on a large-scale dataset of 20 million images; our approach consistently outperforms them.

8.3. Comparison with Inpainting-like Baselines

In the field of diffusion-based inpainting, mask-guided mechanisms are widely utilized. To validate our approach, we directly compare it with a Inpainting-like baseline that adopts the input and noise conditioning scheme from inpainting methods. Specifically, this baseline modifies the backbone to a 9-channel input, where the LR image, noise, and mask (RTDM in our method) are concatenated before being fed into the diffusion model. As shown in Table 11, our proposed method achieves superior performance across all metrics. The results indicate that a straightforward adaptation of such inpainting techniques to RSISR yields suboptimal performance.

8.4. Additional Ablation Studies

Impact of TADL Adjustable Weight. To validate the necessity of the TADL adjustable weight and to investigate its influence, we conduct an ablation study on the hyperparameter α in Eq.6. The objective is to quantify how the strength of texture-rich regions emphasis affects the model's performance. We ablate α by varying its value over a range of settings (e.g., $\alpha = 0, 0.5, 1$) and quantitatively evaluate the resulting models on AID datasets. As shown in Table 12, setting $\alpha = 1$ yields leading performance across all metrics.

Table 15. Ablations of different thresholds on LoveDA-Val.

Training	Inference	PSNR↑	LPIPS↓	DISTS↓
0.32-0.37	0.32	**26.22**	0.3554	0.1934
	0.37	26.04	**0.3533**	**0.1896**
0.35-0.40	0.35	**26.22**	0.3665	0.1992
	0.37	26.16	0.3662	0.1980
	0.40	26.06	0.3659	0.1970
0.38-0.43	0.38	26.20	0.3551	0.1925
	0.43	26.03	0.3580	0.1923

Impact of Texture-aware Sampling Schedules over Different Intervals. Our texture-aware sampling schedule is applied only within a specific range of sampling steps. We evaluate several different such intervals in Table 13. When the interval is set to $[100, 500]$, both PSNR and LPIPS achieve their best values. Overall, the results indicate that activating this strategy in the mid-to-late stages of the sampling process yields superior performance, suggesting that it helps reduce the generation of redundant details in the later steps.

Impact of Textual Descriptions. During inference, using prompts generated by multimodal large language models (MLLMs) introduces substantial latency. For example, reconstructing a 1024×1024 image takes only 9 seconds, whereas generating its textual description with LLAVA [10] requires 27.6 seconds. To avoid this overhead, we therefore use an empty text prompt at inference time. Table 14 compares different prompting strategies, showing that textual prompts do not yield performance improvements. We hypothesize that MLLMs struggle to accurately and comprehensively recognize land-cover content in heavily degraded low-resolution remote sensing images, and thus the textual descriptions they generate offer only limited, and sometimes noisy, guidance to the model.

Impact of RTDM Binarization Threshold across Datasets. In the main text, we report quantitative results on the AID [21] dataset under different training threshold intervals and inference thresholds, from which an optimal interval can be identified. Table 15 further presents the corresponding results on the LoveDA [17] dataset. As can be observed, the optimal threshold on AID differs from that on LoveDA dataset, mainly due to the substantial distribution gap between the two datasets. The two datasets differ substantially in spatial resolution and in the composition of their primary land-cover categories. Nevertheless, as shown in main text Table 2, even with a suboptimal threshold, our method consistently outperforms competing approaches on all evaluated datasets in terms of LPIPS and DISTS. If the

Figure 7. Here are some examples that illustrate the mismatch between no-reference metrics and visual quality. Our method produces higher-quality details, yet obtains worse no-reference scores.

method is tailored to a specific type of sensor data, hyperparameters can be further fine-tuned to achieve better results.

8.5. Limitations of No-Reference Metrics

While numerous no-reference image quality assessment (NR-IQA) metrics have been developed for natural images, to the best of our knowledge, there are currently no widely adopted no-reference metrics specifically designed for RSISR. As a result, we are compelled to adopt well-established natural image NR-IQA metrics—NIQE [13], BRISQUE [14], and CLIP-IQA+ [18]—as auxiliary evaluation metrics. Specifically, NIQE and BRISQUE rely on natural scene statistics, while CLIP-IQA+ is trained on image quality assessment (IQA) datasets of natural images. However, since these metrics were originally designed and calibrated for natural images, they suffer from a significant domain gap when applied to remote sensing images. As illustrated in Figure 7, we observe a clear discrepancy between the numerical scores and the perceived visual quality. We suggest that, for evaluating RSISR tasks, greater emphasis should be placed on reference-based metrics and qualitative analysis. We also look forward to the development of no-reference image quality metrics specifically tailored for remote sensing images.

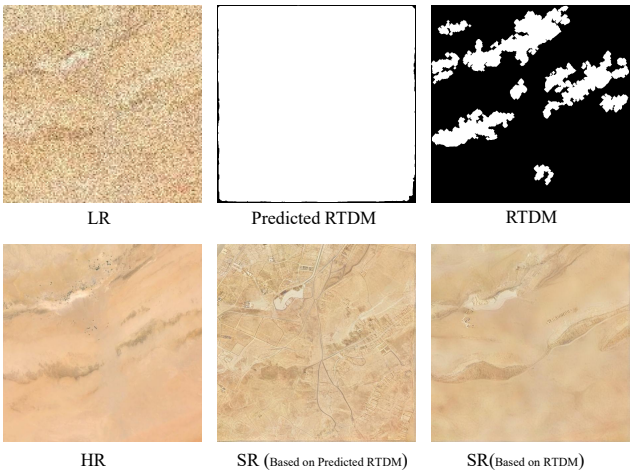

Figure 8. A failure case caused by incorrect RTDM prediction.

9. Limitation

Despite the impressive performance of TexADiff in RSISR, certain limitations remain, primarily reflected in the room for improvement regarding RTDM prediction accuracy. During inference, the model is required to predict the RTDM directly from LR images, which presents a significant challenge when the LR images are accompanied by severe degradation. As shown in Figure 8, prediction deviations in the RTDM lead to anomalous textures in the generated results; in contrast, results based on RTDM estimated from HR images are free of such issues. In practical applications seeking optimal reconstruction quality, human-in-the-loop interaction could be introduced to fine-tune and correct the RTDM. Future research could proceed in two directions: first, exploring a degradation pipeline better tailored to the characteristics of remote sensing images, as the currently adopted Real-ESRGAN [19] pipeline is overly aggressive for this domain; second, integrating Vision-Language Models (VLMs) to leverage their robust high-level semantic understanding for more accurate and reliable identification and localization of texture-rich regions.

10. More Visual Comparisons

We present more visual comparisons in Figure 9 (synthetic scenario) and Figure 10 (real-world scenario). The tag "official" indicates that the corresponding result was obtained with the publicly released checkpoint of each baseline method.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 12

[2] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithd-

iff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, 2025. 12

[3] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 12

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 12

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 12

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 11

[8] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 12

[9] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Trans. Graph.*, 44(6), 2025. 14

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 14

[11] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 12

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[13] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 2012. 15

[14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 2012. 15

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 12

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 11

[17] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 14

[18] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 15

[19] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021. 12, 15

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 11

[21] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE TGRS*, 2017. 14

[22] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 14

[23] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE TPAMI*, 2025. 12

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 11, 12

[25] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE TGRS*, 2024. 12

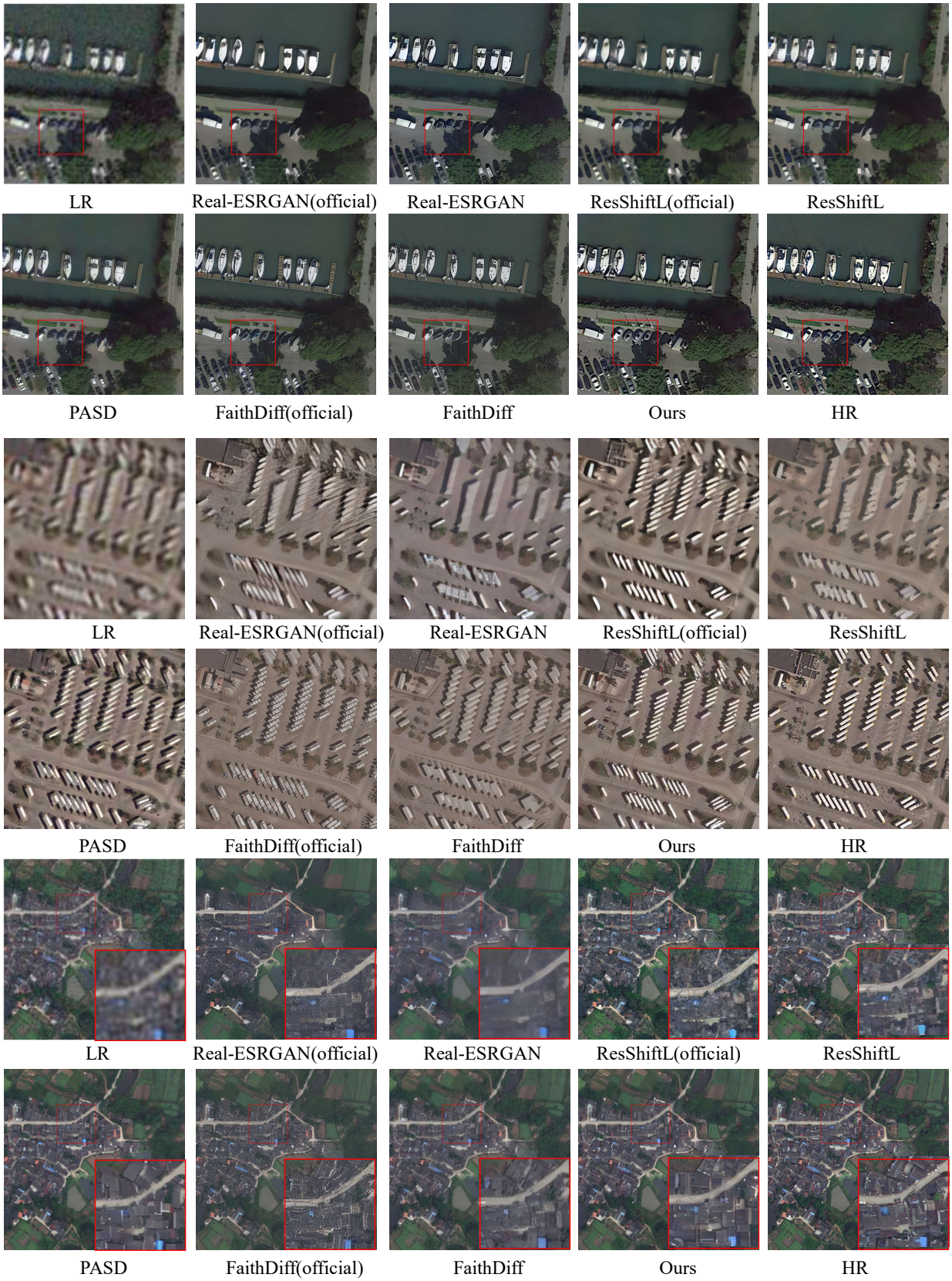

Figure 9. Image SR results (×4) on the synthetic scenario.

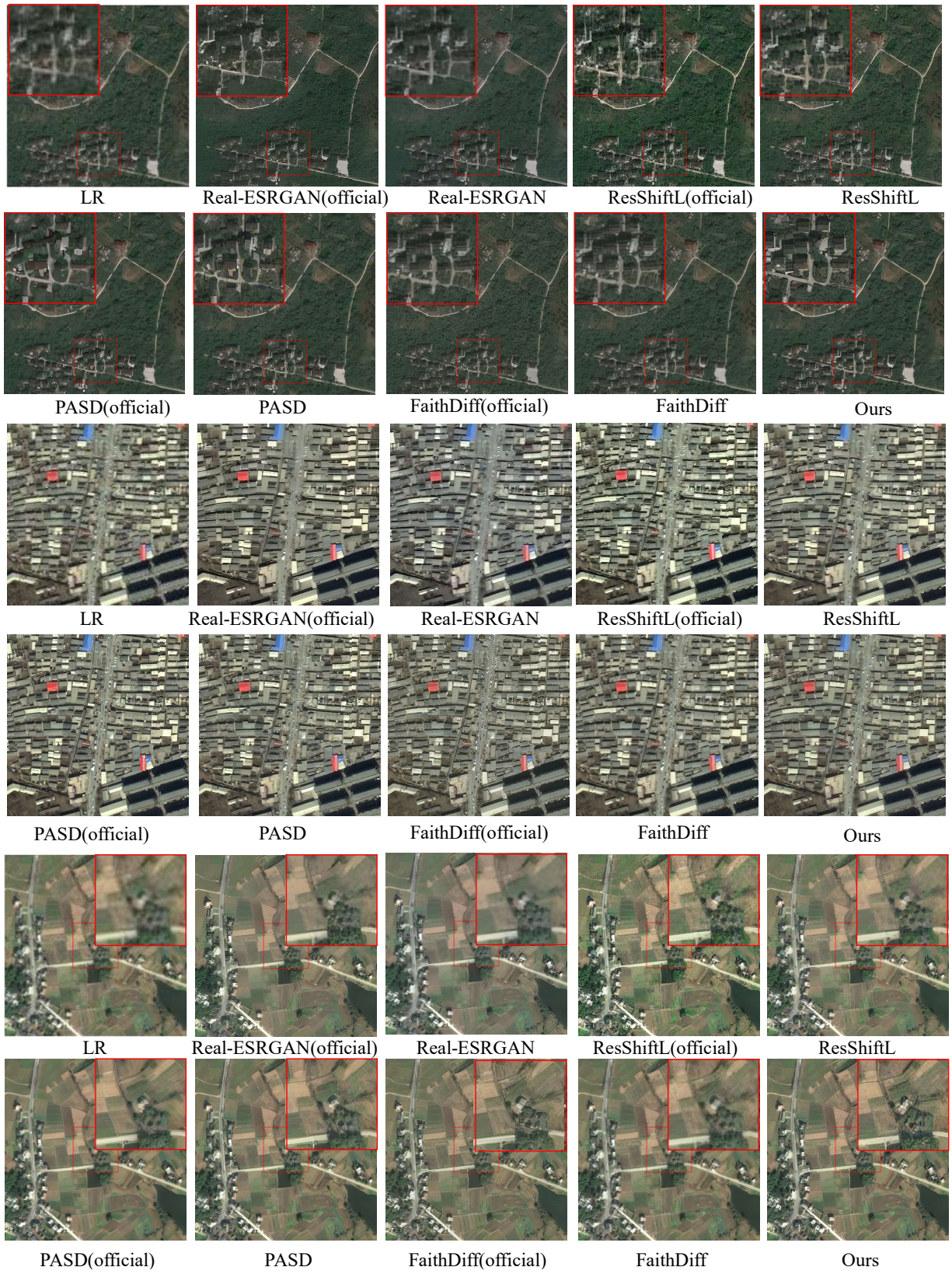

Figure 10. Image SR results (×4) on the real-world scenario.